Data Science Vocabulary

Marketing Definitions

A/B Testing
This is the marketing version of controlled trial experiments. We choose a sample (preferably randomly) to be a control group (A), and another to be a test group (B), and we administer a test by introducing some change to group B, usually a marketing intervention or new product; and we observe the differences between the groups after the change.

Addressable Marketing
Marketing that can be targeted to an individual or household, and can sometimes be connected to subsequent shopping behavior.

Customer Relations Management (CRM)
Processes and Systems that allow manufacturers and retailers to connect directly to some subset of their customers. These systems include loyalty programs and internet manufacturer coupons that require entering an email and generally result in a database of customer-specific information.

NPD’s Checkout
An NPD-proprietary consumer-level purchase-tracking database, populated by a panel of consumers who share receipts.

Statistics Definitions

Cross-validation
Deliberately withholding a subset of the data for the testing of a statistical model. Cross-validation is a good way to produce a model that generates better predictions.

Data Hygiene
Processes that ensure high data quality.
**Data Quality**
The correctness and reliability of a data set.

**Predictive vs. Inferential**
Statistical models built to either predict a future outcome versus models built to explain past behavior. An inferential model is built to be well understood and explainable while a predictive model will generally be more difficult to explain the inner workings of, but will be highly accurate when forecasting. A common example would be that OLS regression is (generally) inferential while a neural network is predictive.

**Sample Bias**
The property of a sample being non-representative of the larger population because some members of the population are less likely to be included than others. For example, if the sample shoppers bought milk or watched TV more than the whole population.

**Computer Programming Definitions**

**Artificial Intelligence**
Intelligence displayed by machines, as opposed to humans. Ideally, this is the ability of machines to ‘perceive’ their environment and ‘succeed’ at some task.

**Amazon Web Service (AWS)**
Amazon Web Service (AWS) is a cloud service for data processing, database storage, hosting, and other large scale processes.

**Big Data**
Data with so many observations or (on rare occasions) so many variables as to be difficult to process using conventional machines and statistical tools.

**Cloud Storage**
Pooled, rentable data storage available through a web source, giving ubiquitous access to information. Users typically have no control over or even knowledge of the physical servers holding their data.

**Cloud Computing**
The processing of data using servers available through some hosted Internet source, giving ubiquitous access to inexpensive processors. Users can typically specify what type of server, and pay only for time they use.
Cluster Computing
Parallel processing of data across groups of (usually cloud) servers; this involves sending of instructions and subsets of data to each of the servers and collecting the results. Typically uses cloud servers. Some languages make this process easy, see Spark, Pyspark.

ETL – Extract, Transform, Load
Data-handling systems primarily developed to access understandable information from large or diverse datasets. Extract pulls the data from its original database(s); transform can do any number of transformations from joining to other data to complicated statistical analyses; load adds the transformed data to a final database, generally from which it can be easily served to a user.

Hadoop
An open source framework for processing and storing extremely large data sets by using distributed computing. Part of the Apache Software Foundation along with Spark and Parquet.

Integrated Development Environment (IDE)
Interactive software used to write code that often includes features useful for debugging, version control, and running code. Common examples include RStudio for R, Atom for Python, and IntelliJ for numerous languages. Environments are not required to write or run code, but they make it easier and faster.

Machine Learning
One type of Artificial Intelligence. Any algorithm that produces results not explicitly written into the program by “learning;” that is, increasing accuracy. An example: a computer program that learned to make new paint colors after reading data from paint chips.

Notebook Technology (Jupyter, iPython)
An interactive type of IDE where individual lines or blocks of code can be run one at a time so that the user can test and display the results of small portions of code.

Open-source
Systems, applications, and programming languages developed collaboratively by people working for free, with the goal of free and open access for all.

Parquet
A highly efficient, column-oriented, data file structure; the big data version of an Excel file.
Python
A general-use programming language that can be used for analytics. Is particularly good for automation. Data scientists will often refer to Pandas, Numpy, and SciPy which are libraries inside of Python.

PySpark
A Python interface for Spark. Other alternatives exist such as SparkR which allows Spark to be accessed using the R programming language.

R
An open source scripted programming language developed to build complex statistical models, data transformations, data visualizations, and analytics. Used to write Shiny Apps.

Remote Application Programming Interface (API or Remote API)
Programming interfaces that allow users to interact with a programming interface through internet protocols. A typical API serves data through a series of URLs, but APIs can do much more than this.

SAS
One of the original statistical programming languages. Handles larger data sets fairly well, has proprietary library files, has a separate macro language which enables some automation. Available through relatively expensive, per-person, per year licenses.

Spark
Apache Spark is an open-source cluster-computing framework. Spark is not a programming language itself and is generally accessed using PySpark, SparkR, Java, or Scala. Spark is important for big data applications since it distributes data across a cluster, allowing significantly faster processing.

SQL
Structured Query Language is used to access and manipulate data from relational databases.

Version control / Source control
A system for sharing code, this manages updates to a code body to ensure the latest versions are being used and the pieces seamlessly work together. Particularly important for large projects that may have hundreds of code files. Github is one of the most popular examples of source and version control.